CUDA

Instructor
Dr. B Krishna Priya

Outline

- The Age of Parallel Processing
- Central Processing Units
- The Rise of GPU Computing
- A brief history of GPUs
- Early GPU computing, CUDA: What is CUDA architecture
- using the CUDA architecture
- Applications of CUDA
  - Medical Imaging
  - Computational Fluid Dynamics
  - Environmental Science

Agile of Parallel Processing

- 2010- Computers are shipped with multicore central processor. Example:
  - Dual-core
  - low-end netbook machines to 8- and 16-core

    workstation computers- supercomputer/ mainframe
  - Portable Music Player
  - Mobile Phone

Contd..

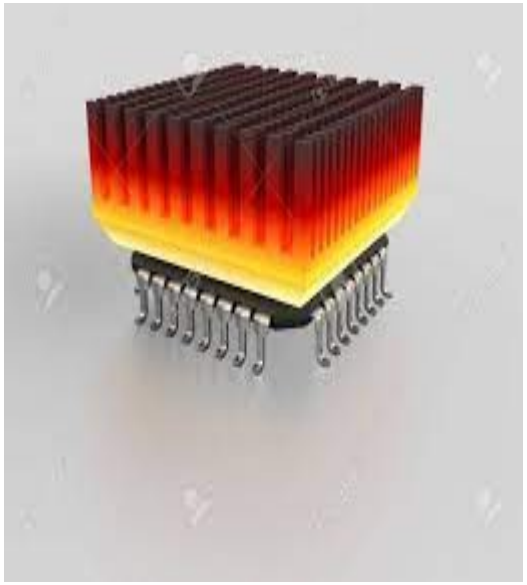
- Software developers will need to cope with a variety of parallel computing platforms and technologies in order to provide novel and rich experiences for an increasingly sophisticated base of users
- Command prompts are out and multithreaded graphical interfaces are in.
- Cellular phones that only make calls are out and phones that can simultaneously play music, browse the Web, and provide GPS services are in.

Central Processing Unit

- 1980s, consumer central processing units (CPUs) ran with internal clocks operating around 1MHz.
- 30 years later- most desktop processors have clock speeds between 1GHz and 4GHz, nearly 1,000 times faster than the clock on the processing original personal computer.
- Increasing the CPU clock speed -performance has been improved, it has always been a reliable source for improved performance.

Contd..

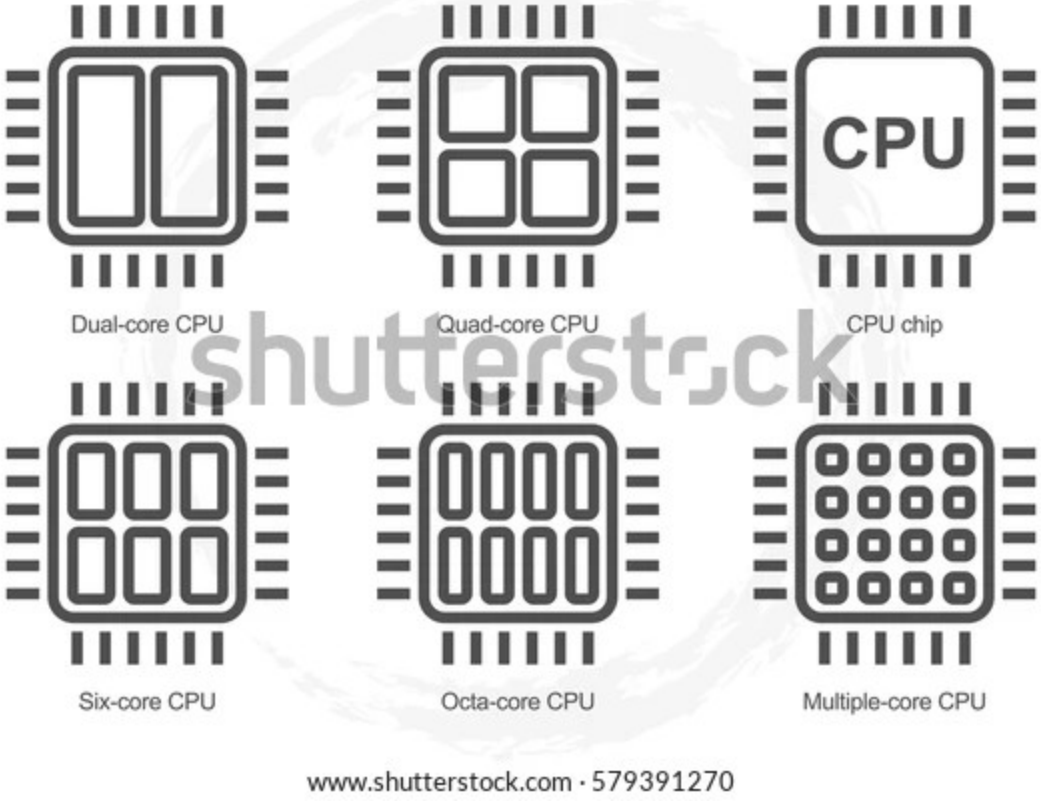
- limitations
  - Limitations of fabrication of integrated circuits.
  - Power and heat restrictions as well as a rapidly approaching physical limit to transistor size.



- Overcome the Limitations

  Personal Computers- three-, four-, six-, and eight-core central processor units

Contd..


CPU
Dual-core CPU
Quad-core CPU
CPU chip
Six-core CPU
Octa-core CPU
Multiple-core CPU
shutterstock
www.shutterstock.com · 579391270

The Rise of GPU Computing

A Brief History of GPU

- In the late 1980s and early 1990s, the growth in popularity of graphically driven operating systems such as Microsoft Windows helped create a market for a new type of processor.
- 1990s-2D display accelerators are used for the personal computers. These display accelerators offered hardware-assisted bitmap operations to assist in the display and usability of graphical operating systems.

Contd..

- 1992- Silicon Graphics opened the programming interface to its hardware by releasing the OpenGL library. Silicon Graphics intended OpenGL to be used as a standardized, platform-independent method for writing 3D graphics applications
- 1992s

1. Doom, Duke Nukem 3D, and Quake helped ignite a quest to create progressively more realistic 3D environments for PC gaming.

2.NVIDIA, ATI Technologies, and 3dfx Interactive began releasing graphics accelerators that were affordable computing enough to attract widespread attention.

Contd.

- NVIDIA’s GeForce 256-transform and lighting computations could be performed directly on the graphics processor, thereby enhancing the potential for even more visually interesting applications.
- NVIDIA’s release of the GeForce 3 series was the computing industry’s first chip to implement Microsoft’s then-new DirectX 8.0 standard.

Early GPU Computing

- The GPUs of the early 2000s were designed to produce a color for every pixel on the screen using programmable arithmetic units known as pixel shaders.
- A pixel shader uses its (x,y) position on the screen as well as some additional information to combine various inputs in computing a final color.

- The additional information could be input colors, texture coordinates, or other attributes that would be passed to the shader when it ran.
- The arithmetic being performed on the input colors and textures was completely controlled by the programmer, researchers observed that these input “colors” could actually be any data
- Inputs were actually numerical data signifying something other than color, programmers could then program the pixel shaders to perform arbitrary computations on this data.
- The results would be handed back to the GPU as the final pixel “color,” although the colors would simply be the result of whatever computations the

CUDA Architecture

- The GeForce 8800 GTX was also the first GPU to be built with NVIDIA’s CUDA Architecture.
- The CUDA Architecture included a unified shader pipeline, allowing each and every arithmetic logic unit (ALU) on the chip to be marshaled by a program intending to perform general-purpose computations.
- NVIDIA intended this new family of graphics processors to be used for general-purpose computing.
- The execution units on the GPU were allowed arbitrary read and write access to memory as well as access to a software-managed cache known as shared memory.

Using the CUDA Architecture

- CUDA C became the first language specifically designed by a GPU company to facilitate general-purpose computing on GPUs.
- A specialized hardware driver to exploit the CUDA Architecture’s massive computational power.

Applications of CUDA

- Medical Imaging
- Computational fluids Dynamics
- Environment science